The JAIST-UET-MITI Machine Translation Systems for IWSLT 2015

Hai-Long Trieu[1], *Thanh-Quyen Dang*[2], *Phuong-Thai Nguyen*[2], *Le-Minh Nguyen*[1]

[1]Japan Advanced Institute of Science and Technology
[2]University of Engineering and Technology, Vietnam National University, Hanoi

trieulh@jaist.ac.jp, dangthanhquyen@gmail.com, thainp@vnu.edu.vn, nguyenml@jaist.ac.jp

Abstract

This paper describes the submission of the Japan Advanced Institute of Science and Technology and the University of Engineering and Technology, Vietnam National University, Hanoi for the machine translation track of the IWSLT 2015 workshop. We participated in the shared task for the language pair: English-Vietnamese. First, we investigate and apply some approaches and techniques including phrase-based, syntax-based and domain adaptation for the TED talks domain. Then, we observe and evaluate experimental results of these systems on the development sets to setup the best configurations. Experimental results show that the phrase-based systems obtain the best performance on this domain in comparison with the other applied approaches.

Keywords: phrase-based machine translation, syntax-based machine translation, domain adaptation

1. Introduction

This year's machine translation track of the IWSLT workshop is for language pairs: English paired with French, German, Chinese, Czech, Thai, and Vietnamese. We participate in both translation directions for English-Vietnamese.

We approach the task by first investigating some effective existing methods: phrase-based and syntax-based. Phrase-based translation systems (Koehn et al., 2003 [16], Chiang, 2007 [5]) achieve state-of-the-art results in many typologically diverse language pairs. For this shared task, we participate in translation for English-Vietnamese, a diverse language pair with many different characteristics in linguistic; therefore, we try to apply the syntax-based approach to exploit linguistic knowledge. For the phrase-based methods, we built our systems based on the Moses toolkit (Koehn et al., 2007 [15]). For the syntax-based methods, we applied the open source Joshua [17] with two particular SCFG types: Hiero [5] and Syntax Augmented Machine Translation (SAMT) [34]. In addition to these two methods, because we used unconstrained data in training our models, we conducted experiments on some domain adaptation techniques including: fill-up [2] and back-off [1] to leverage more improvements in our systems. We evaluated our systems on tuning data sets provided by the workshop.

The rest of this paper is organized as follows: in Section 2, we discuss some linguistic characteristics of the diverse language pair English-Vietnamese and review some previous researches of machine translation for English-Vietnamese. In Section 3, we describe a general system overview with details on our training pipeline and decoder configuration. Next we present empirical results for the individual translation directions. In Section 5, we investigate some challenges in the translation task for TED data. We analyze translation errors and experimental results in Section 6. Finally, conclusions are described in Section 7.

2. English-Vietnamese Machine Translation

In this section, we discuss different characteristics between English and Vietnamese. Then, we review some previous researches related to English-Vietnamese machine translation.

2.1. English vs. Vietnamese: Some Linguistic Characteristics

There are many different characteristics between English and Vietnamese languages. For instance, in word order, adjectives follow nouns in Vietnamese while this order is converse in Vietnamese. In another aspect, English uses morphological morphemes to mark tense and number, whereas Vietnamese uses words that precede the verb to mark tense and the addition of numerals and quantifiers for indicating numbers. See Table 5 of [30] for more details of these comparisons.

2.2. Previous Work

Dinh et al., 2003 [8] presented a hybrid model for machine translation (MT) which combines rule-based MT and corpus-based MT (Bitext-Transfer Learning) that learns from bilingual corpus to extract disambiguating rules. Rule-based MT systems were improved by using word-order transfer [9]. This model has been experimented in English-to-Vietnamese MT system (EVT). Ho et al., 2008 [1] built an English-Vietnamese statistical machine translation (SMT) system namely EVSMT1.0 based on the framework of the open

[1] http://www.statmt.org/moses/?n=Advanced.Domain#ntoc3

source Moses and showed potential features in comparison with an existing commercial MT using traditional rule-based approach.

For experiments on the language pair English-Vietnamese, Nguyen and Shimazu 2006 [22] proposed a syntactic transformation model in the pre-processing phase which reorder the structure of source sentence so that it is closer to the structure of target sentence. The transformation is also produced by a dependency-based parser together with a set of hand-crafted rules [13]. Nguyen et al., 2006 used linguistic knowledge of languages in the preprocessing phase using a morphological analysis or POS tagger on the source sentence [21]. Nguyen et al., 2008 [32] proposed reordering at trunk level and incorporate the global reordering model into the decoder. Related to syntactic approaches, Nguyen et al., 2008 [23] applied a tree-to-string phrase-based method which employs a syntax-based reordering model in the decoding phase.

There have been efforts in developing English-Vietnamese bilingual corpora. Nguyen et al., 2006 [31] described dictionaries used in English-Vietnamese Machine Translation (EVMT). Another work of building bilingual corpus was conducted in the National project VLSP (Vietnamese Language and Speech Processing). [2] In this project, an English-Vietnamese bilingual corpus was built, which includes more than 100,000 sentence pairs. English-Vietnamese corpora were also built at different levels including a study on building POS-tagger for bilingual corpora or building a bilingual corpus for word sense disambiguation ([6], [7]). This task was also shown in some other researches ([18], [19], [20]).

3. System Overview

3.1. Pre-processing

We pre-processed English training data by using scripts from the Moses toolkit including tokenization, and then truecasing. For Vietnamese training data, we used JVnTextPro[3] for tokenization. We remove sentences longer than 80 words and their corresponding translations.

3.2. Word Alignment

Word alignment was computed using the first three steps of the train-factored-phrasemodel.perl script packed with Moses (Koehn et al., 2007). We used MGIZA++ (Gao and Vogel, 2008) [11], a multi-threaded implementation of GIZA++ (Och and Ney, 2003) [25] using the *grow-diag-final-and* heuristic (Koehn et al., 2003) [16].

3.3. Language Model

We used all available monolingual data and KenLM [12] to train interpolated Kneser-Ney discounted 5-gram LMs for each system.

3.4. Baseline Features

We follow the standard approach to SMT of scoring translation hypotheses using a weighted linear combination of features. The core features of our models are a 5-gram LM score, phrase translation and lexical translation scores, word and phrase penalties, and a linear distortion score.

We used the hierarchical lexicalized reordering model (Galley and Manning, 2008) [10] with 4 possible orientations (monotone, swap, discontinuous left and discontinuous right) in both left-to-right and right-to-left direction with the setup *msd-bidrectional-fe lexicalized* reordering.

3.5. Tuning and Decoding

The feature weights were tuned using k-best batch MIRA (Cherry and Foster, 2012) [4]. This is a version of MIRA (a margin based classification algorithm) which works within a batch tuning framework. We set the number of inner MIRA loops to 300 passes over the data.

4. Experimental Results

In this section we describe peculiarities of individual systems and present experimental results.

4.1. Data

4.1.1. Bilingual Data

In addition to the data provided by the workshop [4] (constrained data) [3], we used unconstrained data including bilingual corpora for training translation models and monolingual corpora for training language model (LM). Bilingual corpora and sentence length statistics are indicated in Table 1 and Table 2.

The unconstrained bilingual data include several resources in which we used the English-Vietnamese bilingual corpus provided by the National project VLSP (Vietnamese Language and Speech Processing).[5] This corpus includes 80,000 sentence pairs in Economics-Social topics and 20,000 sentence pairs in information technology topic. In addition, we used the EVBCorpus including texts extracted from books, fictions or short stories, law documents, and newspaper articles and then translated by skilled translators [19], [20]. We also used our in-house data including bilingual sentences extracted from newspaper articles. We combines these datasets and obtained 419,385 unconstrained parallel sentences.

For development data, we experimented and evaluated our systems on various tuning sets: each particular set of five development sets (*dev2010*, *tst2010*, *tst2011*, *tst2012*, *tst2013*) provided by the workshop and a set of merging all

[2] http://vlsp.vietlp.org:8080/demo/?page=home

[3] http://jvntextpro.sourceforge.net/

[4] https://wit3.fbk.eu/mt.php?release=2015-01

[5] http://vlsp.vietlp.org:8080/demo/?page=home

these five sets. We setup the development set *tst2013* which shows the best performance for tuning data.

Table 1: Bilingual Corpora. Language codes: en=English, vi=Vietnamese.

Corpus	SentPairs	Tokens en	Tokens vi
Constrained	133,082	54,139	26,867
Unconstrained	419,385	84,506	41,120
Development	1,304	3,918	2,694

Table 2: Sentence Length Statistics. Len.Avg: average sentence length on the corpus. Len.Max: the maximum sentence length. Len>80: number of sentences which length >80.

Corpus	Len.Avg	Len.Max	Len>80
train.en	17.33	513	341
train.vi	22.66	735	1572
test.en	16.54	90	1
test.vi	21.31	120	8

4.1.2. Monolingual Data

For monolingual data, we used English corpora of the WMT 2015,[6] which are permissible in the workshop IWSLT 2015. For Vietnamese data, we crawled articles from *wikipedia* by using more than 1.3B titles provided at *dumps.wikimedia.org*.[7] In addition, we crawled and extracted 800,000 Vietnamese articles from the website *baomoi.com*.[8] These articles were then pre-processed to produce a huge Vietnamese monolingual data. These monolingual data are shown in Table 3.

Table 3: Monolingual Data

Corpora	Sentences	Tokens
en	46,788,513	20,665,762
vi	21,180,758	1,960,909

4.1.3. Test Data

Test data of the workshop IWSLT 2015 include 1080 sentences on both English-Vietnamese and Vietnamese-English extracted from 12 talks of TED data. Statistics of sentences length of the test sets are shown in Table 2. The average length of the English and Vietnamese sets are 16.54 and 21.31, respectively. There are few sentences with length greater than 80.

[6] http://www.statmt.org/wmt15/translation-task.html
[7] http://dumps.wikimedia.org/viwiki/20150901/
[8] http://www.baomoi.com/

4.2. Experiments on Syntax-based Approaches

We found that advantages of syntax-based translation can resolve some differences between English and Vietnamese discussed in Section 2.1 including: i) reordering for syntactic reasons – e.g., move Vietnamese adjectives follow nouns ii) better explanation for function words – e.g., prepositions, determiners iii) conditioning to syntactically related words – translation of verbs may depend on subject or object.

Therefore, in our experiments, we attempted to apply syntax-based methods for the machine translation track. We used Joshua, a Java-based open source implementation of the hierarchical decoder (Li et al., 2009)[17], release 6.0.

Throughout this work, we applied two particular SCFG types known as Hiero (Chiang, 2007) and Syntax Augmented Machine Translation (SAMT) (Zollmann and Venugopal, 2006). We used Thrax (Weese, 2011) [14], an open-source grammar extractor for Hiero and SAMT grammars. We built systems for two language pairs for the IWSLT 2015 shared task: vi-en and en-vi. For the vi-en language pair, we built both SAMT and Hiero grammars, for the en-vi language pair, we only built Hiero grammar.

We used the constrained parallel data to train the translation models. The parallel data was subsampled using Joshua's built in subsampler to select sentences with n-grams relevant to the tuning and test sets. We used SRILM [29] to train a 5-gram LM with Kneser-Ney smoothing using the appropriate side of the parallel data. Before extracting an SCFG with Thrax, we pre-processed the data. For English side, we used the provided Perl scripts to tokenize and normalize the data. For Vietnamese side, we used JvnTextPro to tokenize data. We lower-case data before extracting an SCFG. For SAMT grammar extraction, we parsed the English training data using the Berkeley Parser (Petrov et al., 2006) [27] with the provided Treebank-trained grammar. We tuned the model weights against the tuning sets of the workshop using ZMERT (Zaidan, 2009) [33], an implementation of minimum error-rate training included with Joshua. We decoded the test set to produce a 300-best list of unique translations, then chose the best candidate for each sentence using Minimum Bayes Risk reranking (Kumar and Byrne, 2004) [28]. To re-case the 1-best test set output, we trained a true-case 5-gram LM using the same previous LM training data, and used Perl script to translate from the lowercased to true-case output. Table 4 shows experimental results of the submitted systems (phrase-based) and the syntax-based on the development set.

Table 4: Experimental results on the tuning data (BLEU)

Setup	en-vi	vi-en
SAMT	–	9.91
Hiero	19.27	12.52
Phrase-based (in-domain)	23.92	12.94
Phrase-based (out-of-domain)	25.49	18.27

We use BLEU [26] as the metric to evaluate our systems. Experimental results in Table 4 show higher BLEU scores of the phrase-based compared with the syntax-based methods on the development data. This kind of data, spoken languages, includes complicated structure sentences that we will discuss in Section 5. These characteristics lead to challenges for the syntax-based methods in parsing sentences into syntax structures. Since the results on development data, we set syntax-based outputs as contrastive runs, and phrase-based outputs are submitted to the workshop as primary runs.

4.3. Experiments on Domain Adaptation

Since we used unconstrained bilingual data from other domains in the phrase-based method, we attempted to apply some strategies for domain adaptation including fill-up and back-off combinations. We show experimental results of domain adaptation in this section.

Fill-up Combination (Bisazza et al., 2011 IWSLT): Fillup preserves all the entries and scores coming from the first model, and adds entries from the other models only if new. Moreover, a binary feature is added for each additional table to denote the provenance of an entry. These binary features work as scaling factors that can be tuned directly by MERT [24] along with other models' weights.

Back-Off Combination: This is a simplified version of fill-up. Nevertheless back-off technique does not generate the binary feature denoting the provenance an entry, and this makes the main advantage of back-off: the combined table contains the exact number of scores of their combining tables.

Table 5: Experimental results on domain-adaptation techniques (BLEU). Domain-adaptaion techniques: *fill-up* and *back-off*. *Merged-data*: merging in-domain and out-of-domain data for training.

Setup	en-vi	vi-en
Fill-up	27.90	17.68
Back-off	28.08	17.74
Merged-data	28.32	22.02

We compared results produced by fill-up and back-off techniques with those of the setup *merged-data* in which we merge all in-domain and out-of-domain data together and then train a translation model to obtain only one phrase table. Experimental results in Table 5 show higher scores in the merged-data training setup. Therefore, the merged-data setup was used for generating the primary runs.

4.4. Results

To train translation models, we merged the constrained and unconstrained bilingual corpora, then we run the processing steps described in Section 3. Table 6 shows BLEU scores of our translations on the evaluation system of the workshop.[9]

Table 6: Experimental results on the test sets IWSLT 2015 (BLEU). Hiero, SAMT: syntax-based systems. Submitted system: Phrase-based (out-of-domain).

Setup	en-vi	vi-en
baseline	27.01	24.61
SAMT	–	15.16
Hiero	21.48	15.05
Phrase-based (in-domain)	26.57	16.51
Phrase-based (out-of-domain)	28.17	21.53

We investigated and experimented syntax-based approaches using SAMT and Hiero grammars, which are described in Section 4.3. We used in-domain data for these systems. For English to Vietnamese translation (en-vi translation), Hiero shows a BLEU score of 21.48 which is 5.09 lower than the phrase-based method (26.57). For Vietnamese-English translation (vi-en translation), the result of Hiero is 1.46 lower than that of the phrase-based method (15.05 vs. 16.51). Meanwhile, SAMT which is experimented only on vi-en translation shows a slightly higher score than that of Hiero (15.16 vs. 15.05). For both translation directions, the phrase-based systems show higher BLEU scores than the syntax-based systems. The submitted system, phrase-based (out-of-domain), shows the highest BLEU scores (28.17 for en-vi translation, and 21.53 for vi-en translation). In comparison with the phrase-based in-domain system, the phrase-based out-of-domain system obtains higher BLEU scores (+1.6 for en-vi and +5.02 for vi-en translations) because of the supplemented data.

In comparison of translation directions, all systems show higher BLEU scores in en-vi than vi-en translations. In the result of Hiero, BLEU score of en-vi translation is 6.43 higher than that of vi-en translation. Similarly, the higher BLEU scores are +10.06 (phrase-based in-domain) and +6.64 (phrase-based out-of-domain).

In comparison with the baseline system of the workshop, our en-vi system shows the better result (28.17 vs. 27.01). Nevertheless, our vi-en system is worse than the baseline (21.53 vs. 24.61).

We will discuss these experimental results in the section of error analyses (Section 6).

5. Data Analysis

The data for machine translation track of the IWSLT 2015 are subtitles from TED talks. Since these data are in spoken language, there are some challenges for translation on this kind of data. We discuss several challenges with some examples.

[9] http://iwslt-server.fbk.eu/eval/Eval.html

<title> Rachel Pike: The science behind a climate headline </title>
Recently the headlines looked like this when the Intergovernmental Panel on **Climate Change**, or IPCC, put out their **report** on the state of **understanding of the atmospheric system**.
That report was written by 620 scientists from 40 countries.
They wrote almost a thousand pages on **the topic**.

Figure 1: An example of relationships in contexts and topics between sentences of TED data, emphasis (bold) added by author.

5.1. Context and Topic

The first problem is that there exists a connection between different sentences in a text. Sentences in a TED talk's subtitles may be related to each other in terms of context and topic. As shown in Figure 1, phrases *that report* and *the topic* are mentioned previously, and this can be seen as a dependent relationship between sentences. This kind of data causes the translation task more complicated than that of written texts in general.

5.2. Abstract Meaning

A characteristic of spoken languages like TED data is abstract meaning. As shown in Figure 2, *closet* does not mean *a cupboard or wardrobe*. Speakers sometimes tend to use metaphors in their speech, and it is not easy for machine translation systems to correctly produce output. This is also another challenge in translation tasks for TED data.

<title> Ash Beckham: We're all hiding something. Let's find the courage to open up </title>
<seg id="1"> I think we all have **closets**. </seg>
<seg id="2"> Your **closet** may be telling someone you love her for the first time, or telling someone that you're pregnant, or telling someone you have cancer, or any of the other hard conversations we have throughout our lives. </seg>

Figure 2: An example of abstract meaning in TED data, emphasis (bold) added by author.

5.3. Sentence Structures

Unlike written texts, structures of sentences in TED data are usually quite complicated, and this is a particular characteristic of spoken languages. This is not easy to realize and parse syntactic structures for sentences accurately. This also leads to the applying of syntax-based approaches for this kind of data more difficult. Figure 3 shows an example of this challenge.

<title> Mary Lou Jepsen: Could future devices read images from our brains? </title>
<seg id="14"> But that experience, I think, gave me a new appreciation for men and what they might walk through, and I've gotten along with men a lot better since then. </seg>

Figure 3: An example of complicated sentence structures in TED data.

Table 7: Out Of Vocabulary Statistics (%)

Setup	en-vi	vi-en
SAMT	–	1.67
Hiero	6.90	2.44
Phrase-based (in-domain)	5.03	2.50
Phrase-based (out-of-domain)	2.97	1.34

6. Errors Analysis

6.1. Out Of Vocabulary

We show statistics of out-of-vocabulary (OOV) of our systems on the test sets *tst2015*, which are described in Table 7. This is ratio of vocabulary of the test sets that cannot be translated by our systems to produce hypotheses. For the en-vi translations, that ratio of the Hiero (6.90%) is higher than that of the phrase-base in-domain (5.03%). The lowest OOV ratio is of the phrase-based (out-of-domain) which uses unconstrained data. This is also similar to the case of vi-en translations of the phrase-based (out-of-domain). Nevertheless, in the SAMT for vi-en translation, though the OOV ratio is lower than that of the phrase-based (in-domain) (1.67 % vs. 2.50 %), the SAMT still obtains a lower BLEU score (15.16 vs. 16.51). The systems may produce output phrases that differ from reference phrases even when input phrases are included in phrase tables. We discuss some examples of this problem in Section 6.2.

6.2. Hypotheses and Reordering

In Table 8 and Figure 4 two examples of translations are reported, analyzed in the following. In Table 8, we indicate some problems in vi-en translation in terms of meaning and tenses. The input phrase *được nhìn nhận* is translated into *was seen* (phrase-based), *visible* (syntax-based), *has been viewed* (reference, we use the file input of vi-en translation as the reference for vi-en translations). For another example, the input phrase *có thể suy nghĩ* is translated into *could think* (phrase-based and syntax-based), *can think* (reference). As we previously discussed in Section 2.1, Vietnamese differs from English in that it does not morphologically mark tenses. In this example, the two Vietnamese phrases are translated into results which are different from the reference. Recognizing tenses is a challenge for translation systems. This factor can be seen as a reason why

Table 8: An example of Vietnamese to English translation, bold phrases discussed in Section 6.2.

Input	<title>**Alex Wissner-Gross: A new equation for intelligence**</title> <**seg id="2"**>Nếu chúng ta nhìn lại lịch sử xem trí thông minh **được nhìn nhận** thế nào **ta có thể tham khảo** câu nói nổi tiếng của Edsger Dijsktra: "**Hỏi rằng** liệu máy **có thể suy nghĩ** được hay không cũng **thú vị** như hỏi **liệu một chiếc tàu ngầm có bơi được hay không**."</seg>
Phrase-based Output	If we look back in history to see the intelligence **was seen**, **we can refer to** the famous Edsger Dijsktra: **asking** whether machines **could think** or as **exciting** as the question of **whether a submarine had to swim or not**,
Hiero Output	If we look at what intelligence **visible** like, **we can go** even famous saying edsger_dijsktra history: "**or not asking** if machine **could think** is about as **exciting** as asked **if a submersible swimming or not**. "
Reference	If we take a look back at the history of how intelligence **has been viewed**, **one seminal example** has been Edsger Dijkstra's famous quote that "**the question of** whether a machine **can think** is about as **interesting** as the question of **whether a submarine can swim**."

vi-en translations show a lower performance than that of en-vi translation.

Another problem we would like to discuss in this example is choosing appropriate hypotheses. The input phrase *ta có thể tham khảo* is translated into *we can refer* (phrase-based), or *hỏi rằng* is translated into *asking* (phrase-based); *thú vị* is translated into *exciting* (phrase-based, syntax-based). These hypotheses can be accepted in terms of appropriate meaning. Nevertheless, they may be not matched with results of the reference: *one seminar example*, *the question of*, *interesting*, respectively. Therefore, choosing an appropriate hypothesis is another problem that should be solved to improve the translation performance.

In experimental results shown in Table 4 and Table 6, the syntax-based systems show lower scores than the phrase-based systems. Syntax-based methods may less appropriate for this kind of data, TED talks, than phrase-based methods. Nevertheless, we will show here an example that the syntax-based system produces a better result than the phrase-based in terms of reordering. In Figure 4, we describe translations of an English input sentence in the test set with the reference, phrase-based and syntax-based systems, respectively. The input noun phrase *This tool use ability* is translated by the phrase-based and syntax-based systems with different order in output phrases. The input phrase *use ability* which precedes the verb *will have* is a part of the subject, but its translation produced by the phrase-based system follows the verb and now becomes an object of the verb. This causes an incorrect meaning of the output. Meanwhile, this translation of the syntax-based system matches with the reference due to the syntactic analysis in syntax-based methods.

7. Conclusion

In this work, we have described the submitted system of the JAIST-UET-MITI team for the machine translation track of the IWSLT 2015 workshop. This year, we participated in the shared task for the language pair: English-Vietnamese. We investigated and experimented some approaches including phrase-based, syntax-based and domain adaptation. The submitted system, phrase-based approach, is based on the Moses toolkit, which shows the best results on both development sets and test sets in comparison with applied approaches. Although applying some domain adaptation techniques does not improve our unconstrained systems, we will attempt to deal with this by other strategies to obtain better results.

We have discussed some challenges of machine translation for the data domain of the shared task, subtitles of TED talks. We have also analyzed translation errors in some aspects in both approaches: phrase-based and syntax-based. We plan to deal with these issues in the future work.

8. Acknowledgements

This paper has been supported by the project "Xây dựng hệ thống dịch tự động hỗ trợ việc dịch các tài liệu giữa tiếng Việt và tiếng Nhật nhằm giúp các nhà quản lý và doanh nghiệp Hà Nội tiếp cận và làm việc hiệu quả với thị trường Nhật Bản" funded by Hanoi Department of Science and Technology.

9. References

[1] Ho Tu Bao, Pham Ngoc Khanh, Ha Thanh Le, and Nguyen Thi Phuong Thao. Issues and first development phase of the english-vietnamese translation system evsmt1. 0. 2008.

[2] Arianna Bisazza, Nick Ruiz, Marcello Federico, and FBK-Fondazione Bruno Kessler. Fill-up versus interpolation methods for phrase-based smt adaptation. In *IWSLT*, pages 136–143, 2011.

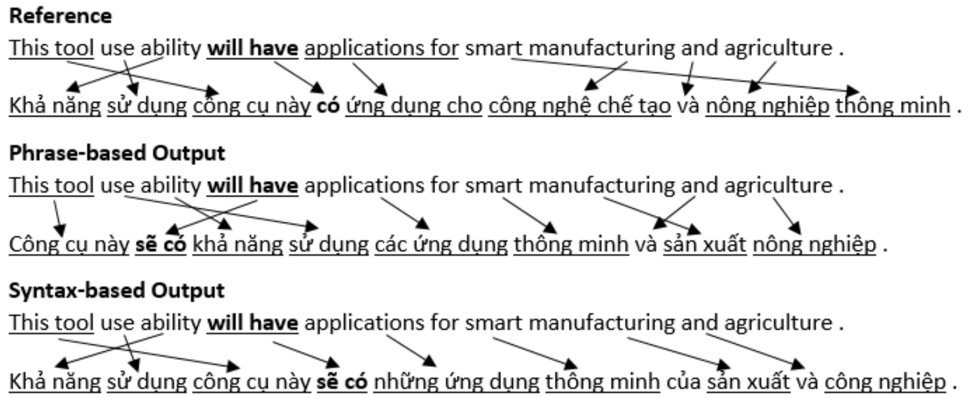

Figure 4: An example of reordering discussed in Section6.2: translations of the reference, phrase-based and syntax-based, respectively. The bold phrases indicate the verbs of the sentences.

[3] Mauro Cettolo, Christian Girardi, and Marcello Federico. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, 2012.

[4] Colin Cherry and George Foster. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436. Association for Computational Linguistics, 2012.

[5] David Chiang. Hierarchical phrase-based translation. *computational linguistics*, 33(2):201–228, 2007.

[6] Dien Dinh. Building a training corpus for word sense disambiguation in english-to-vietnamese machine translation. In *Proceedings of the 2002 COLING workshop on Machine translation in Asia-Volume 16*, pages 1–7. Association for Computational Linguistics, 2002.

[7] Dien Dinh and Hoang Kiem. Pos-tagger for english-vietnamese bilingual corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond-Volume 3*, pages 88–95. Association for Computational Linguistics, 2003.

[8] Dien Dinh, Hoang Kiem, and Eduard Hovy. Btl: a hybrid model for english-vietnamese machine translation. In *Proceedings of the MT Summit IX*, pages 23–27, 2003.

[9] Dien Dinh, Nguyen Luu Thuy Ngan, and Van Chi Nam Do Xuan Quang. A hybrid approach to word order transfer in the english-to-vietnamese machine translation. In *Proceedings of the MT Summit IX*, 2003.

[10] Michel Galley and Christopher D Manning. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 848–856. Association for Computational Linguistics, 2008.

[11] Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57. Association for Computational Linguistics, 2008.

[12] Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics, 2011.

[13] Vu Hoang, Mai Ngo, and Dien Dinh. A dependency-based word reordering approach for statistical machine translation. In *Research, Innovation and Vision for the Future, 2008. RIVF 2008. IEEE International Conference on*, pages 120–127. IEEE, 2008.

[14] Weese Jonathan. A systematic comparison of synchronous contextfree grammars for machine translation. Master's thesis, Johns Hopkins University, 2011.

[15] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[16] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on*

Human Language Technology-Volume 1, pages 48–54. Association for Computational Linguistics, 2003.

[17] Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren NG Thornton, Jonathan Weese, and Omar F Zaidan. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139. Association for Computational Linguistics, 2009.

[18] Quoc Hung Ngo, Dinh Dien, and Werner Winiwarter. Automatic searching for english-vietnamese documents on the internet. In *24th International Conference on Computational Linguistics*, page 211, 2012.

[19] Quoc Hung Ngo and Werner Winiwarter. Building an english-vietnamese bilingual corpus for machine translation. In *Asian Language Processing (IALP), 2012 International Conference on*, pages 157–160. IEEE, 2012.

[20] Quoc Hung Ngo, Werner Winiwarter, and Bartholomäus Wloka. Evbcorpus-a multi-layer english-vietnamese bilingual corpus for studying tasks in comparative linguistics. In *Proceedings of the 11th Workshop on Asian Language Resources (11th ALR within the IJCNLP2013)*, pages 1–9, 2013.

[21] Thai Phuong Nguyen and Akira Shimazu. Improving phrase-based statistical machine translation with morphosyntactic transformation. *Machine Translation*, 20(3):147–166, 2006.

[22] Thai Phuong Nguyen and Akira Shimazu. A syntactic transformation model for statistical machine translation. In *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, pages 63–74. Springer, 2006.

[23] Thai Phuong Nguyen, Akira Shimazu, Tu-Bao Ho, Minh Le Nguyen, and Vinh Van Nguyen. A tree-to-string phrase-based model for statistical machine translation. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 143–150. Association for Computational Linguistics, 2008.

[24] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics, 2003.

[25] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51, 2003.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[27] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 433–440. Association for Computational Linguistics, 2006.

[28] Kumar Shankar and Byrne William. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)*, 2004.

[29] Andreas Stolcke et al. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*, 2002.

[30] Giang Tang. Cross-linguistic analysis of vietnamese and english with implications for vietnamese language acquisition and maintenance in the united states. *Journal of Southeast Asian American Education and Advancement*, 2(1):3, 2007.

[31] Nguyen Chanh Thanh, Nguyen Chi Hieu, Phan Thi Tuoi, et al. Dictionaries for english-vietnamese machine translation. In *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, pages 363–369. Springer, 2006.

[32] Vinh Van Nguyen, Thai Phuong Nguyen, Akira Shimazu, and Minh Le Nguyen. Reordering phrase-based machine translation over chunks. In *Research, Innovation and Vision for the Future, 2008. RIVF 2008. IEEE International Conference on*, pages 114–119. IEEE, 2008.

[33] Omar Zaidan. Z-mert: A fully configurable open source tool for minimum error rate training of machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 91:79–88, 2009.

[34] Andreas Zollmann and Ashish Venugopal. Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 138–141. Association for Computational Linguistics, 2006.